PROCEEDINGS OF THE
HUMAN FACTORS SOCIETY
27th ANNUAL MEETING 1983
NORFOLK VA
10 - 14 OCT 1983

THE EFFECT OF DELAYED REPORT ON SUBJECTIVE RATINGS OF MENTAL WORKLOAD

F. Thomas Eggemeier
WRIGHT STATE UNIVERSITY and SYSTEMS RESEARCH LABORATORIES, INC.
Dayton, Ohio

Mark S. Crabtree
SYSTEMS RESEARCH LABORATORIES, INC.
Dayton, Ohio

Patricia A. LaPointe
WRIGHT STATE UNIVERSITY
Dayton, Ohio

ABSTRACT

Forty-eight subjects performed a short-term memory task with several difficulty levels and provided either immediate or delayed ratings of workload via the Subjective Workload Assessment Technique (SWAT). Mean SWAT ratings did not vary significantly as a function of delayed report, but a substantial number of subjects gave delayed ratings that were discrepant from their immediate ratings. A counterbalancing effect in delayed ratings appears to have been a factor in the failure of the delay effect to reach significance. A secondary objective of this study was to examine the sensitivity of SWAT in a between-subjects design. SWAT ratings varied significantly as a function of task difficulty manipulations, supporting the sensitivity of SWAT to the workload of the conditions used.

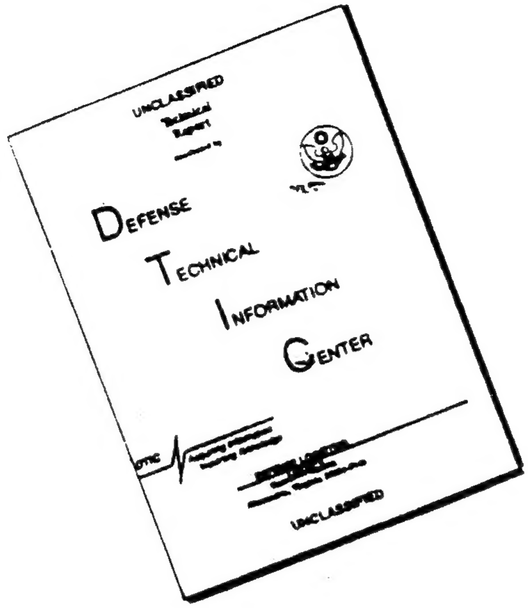
INTRODUCTION

Subjective techniques have been used extensively as measures of operator workload (e.g., Moray, 1982; Williges and Wierwille, 1979). A variety of different techniques (e.g., magnitude estimation, paired comparisons) have been applied in gathering workload judgments, but the rating scale is the most frequently used procedure, especially in simulation or operational environments. The widespread use of rating scales can be attributed to their ease of implementation, lack of intrusiveness on operator performance, and high degree of operator acceptance.

In application-oriented environments, a question exists concerning how the accuracy of subjective ratings might be affected when practical constraints require a delay between task performance and workload estimation. For example, it is frequently maintained that a pilot or operator is too busy during peak workload periods to complete a rating scale, and that workload reports must be delayed until the opportunity arises to complete them. Since subjective ratings depend upon the operator's ability to remember the workload experienced during task performance, delays in rating scale completion constitute retention intervals for the information which is necessary to estimate subjective load. Although the current short-term memory literature (e.g., Klatzky, 1980) clearly indicates that some loss of unrehearsed information will occur at relatively short retention intervals (e.g., 15 to 30 seconds), little data currently exist that address the specific relationship between retention interval and the accuracy of subjective ratings of workload. Therefore, the major purpose of this experiment was to investigate the effect of a short retention interval on subjective ratings of workload.

The procedure used to gather the subjective ratings in this experiment was the Subjective Workload Assessment Technique (SWAT). In SWAT (e.g., Reid, Shingledecker, and Eggemeier, 1981; Reid, Shingledecker, Nygren, and Eggemeier, 1981; Reid, Eggémeier, and Nygren, 1982), subjective workload is defined as being composed of three dimensions: (1) time load, (2) mental effort load, and (3) stress load. Each dimension is represented by an individual three-point rating scale with descriptions for each level of load. SWAT is based on conjoint measurement and scaling (e.g., Krantz and Tversky, 1971; Nygren, 1982) and permits ratings on the three dimensions to be combined into one overall interval scale of workload. In order to identify the appropriate rule for combining the three dimensions into one overall scale, a scale development phase is completed. During this phase, subjects (Ss) rank order the subjective workload associated with the 27 possible combinations that result from the three levels of time, mental effort, and stress load. After completion of scale development, an event scoring phase is initiated. During event scoring, Ss perform the task(s) of interest and rate the time, mental effort, and stress load imposed by task performance. Individual ratings on the three dimensions are then converted to the overall interval scale that was derived during the scale development phase. More detailed discussions of the SWAT procedure can be found in Reid et al. (1981a; 1982).

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

Previous investigations with SWAT have demonstrated that the workload ratings were sensitive to variations in the difficulty of several different tasks, including simulated aircrew radio communications and critical tracking (Reid et al., 1981a), short-term memory (Eggemeier, Crabtree, Zingg, Reid, and Shingledecker, 1982), and probability monitoring (Notestine, 1983). All of these investigations used within-subjects designs to examine SWAT sensitivity. This type of design is appropriate for evaluating SWAT sensitivity, since many applications of workload metrics involve within-subjects designs. For example, the same group of test pilots will frequently participate in all conditions of a display option evaluation conducted in a flight simulator. In some applications, however, practical constraints may make it impossible for the same group to participate in all phases of an evaluation. This raises a methodological question concerning the sensitivity of SWAT to task difficulty differences when a between-subjects design is used. Therefore, a secondary objective of this study was to initially investigate the between-subjects design question as it pertains to SWAT.

Both the delayed rating and between-subjects design issues were addressed by examining SWAT ratings that were completed by Ss after performing a short-term memory update task (e.g., Monty, Taub, and Laughery, 1965). This task required that Ss update and recall the status of several categories of information which changed on a regular basis. The memory update task was chosen for this experiment since previous research had indicated that (1) task difficulty could be effectively varied by manipulating stimulus presentation rate (e.g., Monty et al., 1965); and (2) SWAT was sensitive to such manipulation in a within-subjects design (Eggemeier et al., 1982).

METHOD

Subjects. Ss were 48 introductory psychology students at Wright State University. Ss received extra course credit for their participation in the experiment.

Apparatus. Memory stimulus materials were presented on a 12-inch video monitor which was driven by a Commodore VIC 20 computer. Ss were seated approximately 3 meters from the monitor.

Procedure. Categories of information used in the memory update task were four letters of the alphabet (Q, R, S, T) which appeared individually for 500 msecs on the display. The memory task required that Ss keep track of the number of times that each letter category occurred in a sequence. These sequences averaged 20 individual letters which were distributed across the four categories. At the completion of a letter sequence, recall instructions were presented on the display and Ss completed an answer sheet. Task difficulty was manipulated by varying the rate of letter presentation (interstimulus intervals of 1.0, 2.0, and 3.0 seconds). Presentation rate was a between-subjects variable, with 16 Ss performing the memory task at each rate.

During data collection, a block of three trials was presented to Ss. After each trial, Ss indicated the number of times that each letter category had occurred in the sequence. Absolute error in recalling the number of instances of each category served as the memory performance measure. This was computed by determining the deviation of Ss' response from the correct number for each category, and summing the deviations. The measure was absolute in that no distinction was made between overestimates and underestimates by an S. At the completion of a block of trials, Ss rated the subjective workload by completing separate three-point ratings on time, mental effort, and stress load.

Delay of ratings was a within-subjects variable, so that each S provided a SWAT rating immediately after completion of a block of trials and also after a 15 minute delay period. Order of the delay interval (0 versus 15 minutes) was counterbalanced such that one-half of the Ss completed immediate ratings first, while the other half completed the delayed ratings first. The former Ss performed the memory update task, provided their ratings, and were given a 15 minute rest period. After the rest period, Ss performed a memory update task at the same presentation rate as the first task, played a video game for 15 minutes, and then completed their SWAT ratings for the second memory task. Ss who completed the delayed rating first followed the same procedure in the reverse sequence. Although the presentation rate in both memory tasks was the same for each group of Ss, the actual sequences of letters were different and were counterbalanced across the immediate and delayed rating conditions. One purpose of the 15 minute rest period was to minimize the likelihood that Ss would recognize that the presentation rates of the two tasks were identical. Ss were not informed that a workload rating would be required in the delay condition until the rating was actually requested. The video game that was played during the 15 minute delay required the use of a joystick to maneuver a simulated boat, and was predominantly psychomotor in nature. The game did not specifically require retention of verbal information, and was chosen because it was dissimilar to the memory update task. A dissimilar task was used in order to minimize interference effects and provide a relatively pure estimate of the effects of the 15 minute delay on workload ratings.

Prior to actual data collection, Ss received practice on the memory update task and on performing SWAT ratings. During training, all Ss performed three blocks of training trials with presentation rate/memory category combinations that differed from those used during actual data collection. The combinations used during training included: (1) three categories at a 4.0 second rate, (2) four categories at a 2.5 second rate, and (3) five categories at a 1.0 second rate. It is, therefore, important to note that although actual data collection was conducted under a between-subjects design, all Ss had performed and rated the same group of practice tasks.

During the practice session, Ss also completed the scale development phase of SWAT. Following procedures outlined by Reid et al. (1981a; 1982), interval level SWAT scales with ranges of 0 to 100 were derived for use as the subjective workload measures in subsequent analyses.

RESULTS

Memory performance data were analyzed using a two-factor analysis of variance (ANOVA). Three levels of the presentation rate variable (1.0, 2.0, 3.0 seconds) and two levels of the rating delay variable (0, 15 minutes) were included in the ANOVA. A square root transformation, designed to remove proportional relationships between means and variances that are common in this type of error data, was applied prior to conducting the ANOVA. Figure 1 shows the mean transformed memory error scores as a function of presentation rate and rating delay condition. As is clear from Figure 1, neither presentation rate nor rating delay had a marked effect on performance. The ANOVA confirmed this, and indicated that the main effects of presentation rate [$F(2,45) = 1.20$, $p > .25$], rating delay [$F(1,45) = 0.92$, $p > .25$], and their interaction [$F(2,45) = 0.48$, $p > .25$] were not significant. The nonsignificant effect of rating delay condition was expected, since that factor simply represented whether the workload rating completed subsequent to task performance was immediate or delayed. Likewise, there was no reason to anticipate a significant interaction.

Figure 2 shows mean overall interval SWAT ratings as a function of presentation rate and number of memory categories. A 3 x 2 ANOVA performed on the SWAT data indicated that the main effect of presentation rate [$F(2,45) = 8.14$, $p < .01$] was significant, but that the main effect of rating delay [$F(1,45) = 1.48$, $p < .25$] and the interaction [$F(2,45) = 0.62$, $p > .25$] were not.

Although the rating delay effect was not significant, Figure 2 indicates that there was some tendency for mean immediate and delayed

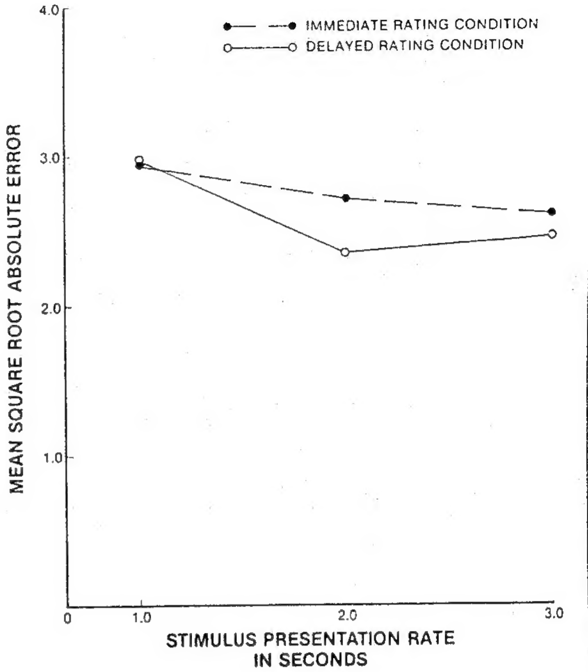


Figure 1. Mean Square Root of Absolute Memory Error as a Function of Stimulus Presentation Rate and SWAT Rating Delay Condition

ratings to differ, particularly in the 2.0 second presentation rate condition. This tendency is supported by the fact that 31 of the 48 Ss assigned ratings under the delay condition that differed from their immediate ratings. Among the Ss who showed a discrepancy between immediate and delayed ratings, 20 Ss increased their ratings in the delayed condition, while 11 Ss decreased their delayed ratings. This trend is reflected in Figure 2, since delayed ratings tend to be higher than immediate ratings. The noted pattern of changes also suggests the existence of a mild counterbalancing effect, where approximately 65 percent of the Ss increased their ratings, while the remainder decreased their ratings. Such a counterbalancing effect represents a potential factor in the lack of a significant delay effect on mean ratings.

Because the presentation rate effect was significant, a Newman-Keuls multiple comparisons test was performed in order to specify the locus of the significant effect(s). This test indicated that SWAT ratings in the 3.0 second condition differed from those in the 2.0 second ($p < .05$) and the 1.0 second ($p < .01$) conditions. The difference between

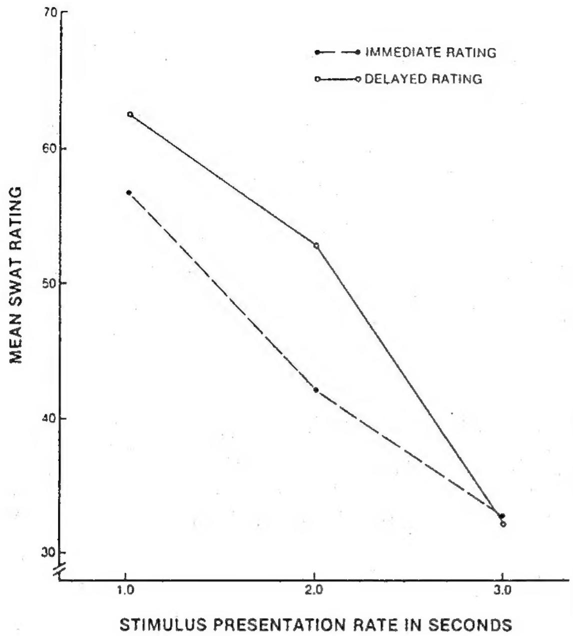


Figure 2. Mean SWAT Ratings as a Function of Stimulus Presentation Rate and Rating Delay Condition

the 1.0 and 2.0 second ratings approached, but did not reach significance. Therefore, in the present between-subjects design, SWAT ratings demonstrated differences in the workload associated with different presentation rates, even though the primary task measure of memory errors did not.

DISCUSSION

The results indicate that mean SWAT ratings were not significantly influenced by the 15 minute delay interval used in this experiment. However, a substantial number of Ss did give delayed ratings that differed from their immediate ones, suggesting that the delay did contribute to some changes in ratings. It is probable that the noted counterbalancing effect of increases versus decreases in the delayed ratings was a factor in the failure to find a significant difference between the mean ratings in the two conditions. If it is assumed that loss of information from short-term memory was a major contributor to the individual discrepancies between immediate and delayed ratings that did occur, there is no reason to expect that a bias in favor of either increases or decreases should have been present in the data. If information lost from memory during the 15 minute delay made it necessary for Ss to guess or estimate the particular levels of load that had been experienced, it appears probable that some Ss would increase their ratings relative to the immediate rating baseline, while others would decrease their ratings. The trend toward a counterbalancing effect that was apparent in the data can, therefore, be interpreted as consistent with a loss from short-term memory of information that is necessary to complete subjective ratings. When such a counterbalancing effect is present, it appears that delays should not significantly alter mean ratings.

The present results are also consistent with those of Notestine (1983), who recently compared immediate and delayed SWAT ratings resulting from performance of a probability monitoring task. Delays of 15 and 30 minutes had no significant effect on mean SWAT ratings, but a number of Ss showed substantial discrepancies between immediate and delayed ratings. The results of the Notestine and the current experiment are, therefore, quite similar.

In applying the results, however, it is very important to note that both studies were specifically designed to test the effects of delays on workload ratings. In each case, a video game that was chosen to minimize interference effects was performed by Ss during the delay interval. As noted earlier, one reason for delaying workload ratings in some applications is the fact that the operator is too busy with continuing or subsequent task performance to complete the necessary ratings. An important area for additional research, therefore, is to investigate the effects of similar intervening tasks on delayed ratings. It is well established in the human memory literature (e.g., Klatzky, 1980) that such retroactive interference effects are an important determinant of forgetting, and that the degree of interference experienced in verbal memory can be related to the similarity of the remembered and intervening material. Since neither the present study nor the Notestine experiment was designed to address the retroactive interference issue, the results should not be generalized to those instances where it is possible that such effects may be present. It is quite possible that such interference effects could introduce a systematic bias into the ratings, destroy any counterbalancing effect, and significantly influence mean ratings. It is also important to note that the current results pertain only to SWAT. It is possible that other rating scale formats (e.g., 10-point scale) may be more or less resistant to the effects of delay than SWAT, which requires that Ss assign relatively simple three-point ratings on the dimensions of time, mental effort, and stress load.

A secondary objective of this study was to initially examine the sensitivity of SWAT in a between-subjects design. The results indicated that SWAT was sensitive to variations in presentation rate in the memory update tasks, and that the ratings were more sensitive to such variations than the primary task measure of memory error. This type of result would be expected from a sensitive measure of workload, since primary task measures such as memory error are generally thought to discriminate overload from nonoverload conditions (e.g., Williges and Wierwille, 1979). Apparently, the difficulty manipulations used in present study were in a nonoverload region, leading to a lack of sensitivity of the primary task measure. The more sensitive subjective technique, on the other hand, successfully discriminated several levels of the variations in load that were employed. These results, which compare favorably with data from previous within-subjects work with the same task (Eggemeier et al., 1982), support the conclusion that SWAT can be a sensitive workload index in a between-subjects design. Although encouraging in this respect, the present results were obtained with pretraining by all Ss on a common set of task difficulty levels, and with a relatively large number of Ss. In spite of the fact that common pretraining of Ss may be possible in operational applications of between-subjects designs, an important topic for future research deals with the effectiveness of between-subjects designs when common training is not provided. A direct comparison of within- and between-subject SWAT sensitivity in the same task difficulty conditions also represents an area for future research.

ACKNOWLEDGEMENT

We wish to thank Mr. Gary B. Reid, Mr. William A. Acton, and 1st Lt. Lee Penick for helpful comments and support with the SWAT data analysis.

REFERENCES

Eggemeier, F. T., Crabtree, M. S., Zingg, J. J., Reid, G. B., and Shingledecker, C. A. Subjective workload assessment in a memory update task. Proceedings of the 1982 Human Factors Society Annual Meeting, October 1982, 643-647.

Klatzky, R. L. Human Memory: Structures and Processes. San Francisco, California: W. H. Freeman and Company, 1980.

Krantz, D. H. and Tversky, A. Conjoint measurement analysis of composition rules in psychology. Psychological Review, 1971, 78, 151-169.

Monty, R. A., Taub, H. A., and Laughery, K. R. Keeping track of sequential events: Effects of rate, categories, and trial length. Journal of Experimental Psychology, 1965, 69, 224-229.

Moray, N. Subjective mental workload. Human Factors, 1982, 24, 25-40.

Notestine, J. The effects of delays in reporting subjective workload ratings: The subjective workload assessment technique in a probability monitoring task. Unpublished Master's Thesis, Wright State University, 1983.

Nygren, T. E. Conjoint measurement and conjoint scaling: A users guide. Wright-Patterson Air Force Base, Ohio: Air Force Aerospace Medical Research Laboratory Technical Report, AFAMRL-TR-82-22, April 1982.

Reid, G. B., Shingledecker, C. A., and Eggemeier, F. T. Application of conjoint measurement to workload scale development. Proceedings of the 1981 Human Factors Society Annual Meeting, October 1981(a), 522-526.

Reid, G. B., Shingledecker, C. A., Nygren, T. E., and Eggemeier, F. T. Development of multidimensional subjective measures of workload. Proceedings of the 1981 IEEE International Conference on Cybernetics and Society, 1981(b), 403-406.

Reid, G. B., Eggemeier, F. T., and Nygren, T. E. An individual differences approach to SWAT scale development. Proceedings of the 1982 Human Factors Society Annual Meeting, October 1982, 639-642.

Williges, R. C. and Wierwille, W. W. Behavioral measures of aircrew mental workload. Human Factors, 1979, 21, 549-574.